

**Sztuczna inteligencja : bezpieczeństwo i zabezpieczenia / redakcja  
Roman V. Yampolskiy ; Bill Joy [i 10 pozostałych]. – Wydanie I. –  
Warszawa, 2020**

Spis treści

<b>Wstęp: wprowadzenie do bezpieczeństwa i ochrony sztucznej inteligencji</b>	<b>xi</b>
<b>Podziękowania</b>	<b>xxvii</b>
<b>Redaktor naukowy</b>	<b>xxix</b>
<b>Współpracownicy</b>	<b>xxxii</b>
<b>Część I Obawy luminarzy</b>	
Rozdział 1 Dlaczego przyszłość nas nie potrzebuje <i>Bill Joy</i>	3
Rozdział 2 Głęboko przeplatana obietnica i niebezpieczeństwo GNR <i>Ray Kurzweil</i>	25
Rozdział 3 Podstawowe pobudki SI <i>Stephen M. Omohundro</i>	59
Rozdział 4 Etyka sztucznej inteligencji <i>Nick Bostrom i Eliezer Yudkowsky</i>	71
Rozdział 5 Przyjazna sztuczna inteligencja: Wyzwanie fizyki <i>Max Tegmark</i>	87
Rozdział 6 MDL destylacja inteligencji: Poznawanie strategii bezpiecznego dostępu do superinteligentnych możliwości rozwiązywania problemów <i>K. Eric Drexler</i>	93
Rozdział 7 Problem uczenia się wartości <i>Nate Soares</i>	111
Rozdział 8 Przykłady kontradyktoryjne w świecie fizycznym <i>Alexey Kurakin, Ian J. Goodfellow i Samy Bengio</i>	123
Rozdział 9 W jaki sposób może zaistnieć SI? Różne podejścia i ich implikacje dla życia we wszechświecie <i>David Brin</i>	141

Rozdział 10 Przyszłość MADCOM: Jak sztuczna inteligencja może wzmocnić propagandę obliczeniową, przeprogramować ludzką kulturę oraz zagrozić demokracji... i co można z tym zrobić <i>Matt Chessen</i>	159
Rozdział 11 Strategiczne implikacje otwartości w rozwoju sztucznej inteligencji <i>Nick Bostrom</i>	183
<b>Część II Odpowiedzi naukowców</b>	
Rozdział 12 Korzystanie z ludzkiej historii, psychologii i biologii w celu uczynienia SI bezpieczną dla ludzi <i>Gus Bekdash</i>	211
Rozdział 13 Bezpieczeństwo SI z perspektywy pierwszej osoby <i>Edward Frenkel</i>	251
Rozdział 14 Strategie dla nieprzyjaznej wyroczni SI z przyciskiem resetowania <i>Ile Häggström</i>	260
Rozdział 15 Zmiany celu w inteligentnych agentach <i>Seth Herd, Stephen J. Read, Randall O 'Reilly i David J. Jilk</i>	273
Rozdział 16 Ograniczenia weryfikacji i walidacji zachowań agencyjnych <i>David J. Jilk</i>	283
Rozdział 17 Kontraduktoryjne uczenie maszynowe <i>Phillip Kuznetsov, Riley Edmunds, Ted Xiao, Humza Iqbal, Raul Puri, Noah Golmant i Shannon Shih</i>	295
Rozdział 18 Uzgadnianie wartości wykorzystując obliczalną odległość preferencji <i>Andrea Loreggia, Nicholas Mattei, Francesca Rossi i K. Brent Venable</i>	313
Rozdział 19 Racjonalnie uzależniona sztuczna superinteligencja <i>James D. Miller</i>	329
Rozdział 20 Bezpieczeństwo aplikacji robotów z wykorzystaniem ROS <i>David Portugal, Miguel A. Santos, Samuel Pereira i Micael S. Couceiro</i>	341
Rozdział 21 Wybór preferencji społecznej i problem wyrównania wartości <i>Mahendra Prasad</i>	363

Rozdział 22 Rozłączne scenariusze katastrofalnego ryzyka SI <i>Kaj Sotala</i>	395
Rozdział 23 Realizm ofensywny i niezabezpieczona struktura systemu międzynarodowego: Sztuczna inteligencja i globalna hegemonia <i>Maurizio Tinnirello</i>	423
Rozdział 24 Superinteligencja i przyszłość rządów: Priorytetyzacja problemu kontroli na końcu historii <i>Phil Torres</i>	445
Rozdział 25 Wojskowa SI jako zbieżny cel samodoskonalącej się SI <i>Alexey Turchin i David Denkenberger</i>	467
Rozdział 26 Wrażliwe na wartości podejście do projektowania inteligentnych agentów <i>Steven Umbrello i Angelo F. De Bellis</i>	491
Rozdział 27 Konsekwencjalizm, deontologia i bezpieczeństwo sztucznej inteligencji <i>Mark Walker</i>	509
Rozdział 28 Inteligentne maszyny są zagrożeniem dla ludzkości <i>Kevin Warwick</i>	523
<b>Indeks</b>	<b>533</b>

oprac. BPK